Assignment III & IV: Deep Learning and Transfer Learning

July 14, 2025

ID5003W: Industrial AI at Scale Laboratory

Total Marks:100

Submission Deadline : August 5, 2025

Objective

Use the CIFAR-10 dataset (available via torchvision.datasets) to build two different type of classifiers to classify objects given in the input images.

- Feature Extraction + Spark ML Classification: Use the class example code to extract features and apply Spark ML classifiers to build classification models.
- CNN Model Training Using BigDL: Train a CNN model from scratch using BigDL.

Use only 10,000 images from the training dataset and 500 images from the test dataset for training and evaluation.

Part I : Data Analysis of the data (2 Marks)

1. Plot the class distribution in the complete train and test datasets.
2. Plot the class distribution in the selected subset of the train and test datasets.

Use google colab for the plots.

Part II : Feature Extraction and Classification (45 Marks)

1. Use the class example code to extract features from the CIFAR-10 dataset.
2. List the available classifiers in the Spark ML library that can be used to build classifiers using the extracted features.
3. Select three models among them and justify your selection.
4. Train the selected models and compare their results using suitable evaluation metrics.

5. You are allowed to use a maximum of 8 total cores and 10 GB executor memory for this assignment. Try three different parallelism settings using different combinations of spark.executor.cores, spark.cores.max and spark.executor.memory. Show the results and explain which parallelism configuration performed the best and why.

6. In your code, clearly indicate where parallelism is being used.

Part III : CNN Model Training Using BigDL (45 Marks)

Follow the instructions from the class notes to install BigDL and use Orca for distributed training.

1. Build a CNN model using the same dataset and subset size, with your choice of layers. Justify your design choices.

2. Train the model using three different parallelism settings (using at most 8 cores total).

3. Train the model using 20000 images instead of 10000 using the best spark configuration obtained in question 2.

Part IV : Comparative Analysis (8 Marks)

1. Compare the accuracy and training time of all your trained models.

2. Analyze the results and explain which model performed the best with respect to accuracy, training time, and resource utilization.

Execution Instructions

1. Use Google Colab and the lab cluster for Part I and II.

2. Use your personal PC for Part III.

3. In your Spark code, setting both spark.executor.cores and spark.cores.max is mandatory. Ensure that spark.cores.max is divisible by spark.executor.cores.

4. Use the following command to submit your job to the server:

   ```
   python ./spark_job_execution.py ./your_code.py your_rollno
   ```

5. Check the log file for the output.

Submission Instructions

1. Submit the log files generated during execution.

2. Submit a report in PDF format. (Notebook submissions are not accepted.)

3. Submit the exact code that you submitted to the Spark cluster.